# Methodology and resources for the structural segmentation of music pieces into autonomous and comparable blocks

Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, Emmanuel Vincent

**Publications Internes de l'IRISA**
**ISSN : 2102-6327**
**PI 1979 – Mai 2011**

# METHODOLOGY AND RESOURCES FOR THE STRUCTURAL SEGMENTATION OF MUSIC PIECES INTO AUTONOMOUS AND COMPARABLE BLOCKS

Frédéric BIMBOT[*] , Emmanuel DERUTY[**] , Gabriel SARGENT[***] , Emmanuel VINCENT[****]

**Abstract:** The approach called *decomposition into autonomous and comparable blocks* specifies a methodology for producing music structure annotation by human listeners based on a set of criteria relying on the listening experience of the human annotator. The present article develops further a number of fundamental notions and practical issues, so as to facilitate the usability and the reproducibility of the approach.

We formalize the general methodology as an iterative process which aims at estimating both a *structural metric pattern* and its *realization*, by searching empirically for an optimal compromise describing the organization of the content of the music piece in the most economical way, around a typical timescale.

Based on experimental observations, we detail some practical considerations and we illustrate the method by an extensive case study. We introduce a set of 350 songs for which we are releasing freely the structural annotations to the research community, for examination, discussion and utilization.

**Key-words:** music structure, annotation, MIR

---

## *METHODOLOGIE ET RESSOURCES POUR LA SEGMENTATION STRUCTURELLE DES MORCEAUX DE MUSIQUE EN BLOCS AUTONOMES ET COMPARABLES*

**Résumé :**

*L'approche dite de* décomposition en blocs autonomes comparables *décrit une méthodologie pour l'annotation manuelle de structure musicale. Elle est fondée sur un ensemble de critères faisant appel à l'expérience d'écoute musicale des annotateurs. Cet article approfondit un certain nombre de notions fondamentales et de questions pratiques, afin de faciliter la mise en oeuvre et la reproductibilité de l'approche.*

*La méthodologie de décomposition est formulée en tant que procédé itératif qui vise à estimer simultanément un* patron métrique structurel *et sa* réalisation*, en recherchant empiriquement un compromis optimal permettant de décrire l'organization du contenu du morceau de la façon plus économique possible, autour d'une échelle donnée.*

*Sur la base d'observations expérimentales, nous détaillons quelques considérations pratiques et nous illustrons la méthode par une étude de cas complète. Nous présentons un ensemble de 350 chansons que nous mettons à disposition de la communauté scientifique, afin qu'elles soient examinées, discutées et utilisées.*

**Mots clés :** *structure musicale, annotation, recherche d'information dans les contenus musicaux*

---

[*] CNRS, IRISA (- UMR6074). *frederic.bimbot@irisa.fr*
[**] INRIA, Centre INRIA Rennes - Bretagne Atlantique (temporary consultant). *emmanuel.deruty@gmail.com*
[***] Université de Rennes1, IRISA (- UMR6074). *gabriel.sargent@irisa.fr*
[****] INRIA, Centre INRIA Rennes - Bretagne Atlantique. *emmanuel.vincent@inria.fr*

# 1 INTRODUCTION

Given its numerous applications, the automatic inference of musical structure is a key subject in MIR [1], which has been focusing significant research effort in the past years [2, 3, 4, 5, 6, 7, 8, 9, 10]. It has also triggered several studies [11, 12] and projects [13, 14] supporting this research with the investigation of methodological issues and the collection of annotated data.

In this context, the structural description approach called decomposition into autonomous and comparable blocks was recently introduced [12] in terms of general concepts, inspired from structuralism and generativism. It has been designed to be applicable to a wide range of "conventional" music, including pop music.

The present paper develops further this approach, with the purpose of providing a more practical annotation methodology, so as to facilitate the usability and the reproducibility of the process. We also present the current state of our annotation effort, namely 350 pop songs, for which we are releasing freely the structural annotations to the research community.

# 2 PROBLEM STATEMENT

## 2.1 Levels of musical organisation

It is commonly agreed that the composition and the perception of music pieces rely on simultaneous processes which vary at different timescales. Similarly to [15], we consider the three following levels corresponding to three different ranges of timescales :

- the low-level elements which correspond to fine-grain events such as notes, beats, silences, etc... We call this level the *acoustic level* and its time scale is typically below or around 1 second.
- the mid-level organization of the musical content, based on compositional units such as bars or hyper-bars or on perceptual units such as musical cells and phrases, ranging typically between 1 and 16 seconds. We will refer to this level as the *morpho-syntagmatic* level.
- the high-level structure of the musical piece, which describes the long term regularities and relationships between its successive parts, and which we will call the level of the *semiotic structure*, typically at a time scale around or above 16 seconds.

The figure of section 6 provides an illustration of these three levels. Note that we use the term *semiotic* in a quite restricted scope, (compared for instance to that of Nattiez [16]) as denoting the high-level *symbolic* and *metaphoric* representation of musical content[1].

## 2.2 Semiotic structure

What we consider as the semiotic structure of a music piece is something that may look like :

**A B C D E F B C D E G D E D E H**

thus reflecting :

1. some sort of high-level decomposition/segmentation of the whole piece into a limited number of blocks (here 16 blocks) of comparable size, and
2. some form of similarity or equivalence relationship between blocks bearing identical labels (here, 8 distinct symbols)

Providing a semiotic description for a music piece requires primarily the identification of the most adequate *granularity* (block size and number of blocks) which then conditions the inventory of labels.

From the example below, choosing a finer granularity could lead to a sequence of labels such as :

**AA'BB'CC'DD'EE'FF'BB'CC'DD'EE'GG'DD'EE'DD'EE'HH'**

where any symbol $X$ is systematically followed by symbol $X$, therefore yielding a rather redundant semiotic description.

Conversely, a coarser granularity would require either the uneven grouping of the units into *irregular* segments (*i.e.* of more diverse sizes) :

**A BC DE F BC DE G DE DE H**

[1] We thus avoid the term *semantic*, referring to the musical *meaning* of objects (for instance, chorus, verse, etc) : such a notion falls completely outside the scope of this paper.

or a very misleading representation such as :

**AB CD EF BC DE GD ED EH**

which completely hides the similarities existing between portions of the piece which had identical labels at a lower scale.

This example thus illustrates a simple case where there exist clearly a preferable granularity at which the semiotic level of the music piece can be described with some form of optimal *compromise* between :

- The minimality of the set of labels
- The informativeness of the sequence of labels
- The regularity of the block size

The goal of this work is to present a set of methodological principles for :

1. identifying the most appropriate granularity for describing the semiotic structure, and
2. locating as univocally as possible the corresponding block boundaries.

In this article, the granularity referred to in item 1 is defined as the *structural metric* of the music piece and the actual borders of the segmental units (item 2) as the *realization* of the structural meter.

The proposed process relies on the listening of music pieces, but can be extended to music in written form (scores). However, note that scores may not be available and sometimes are even meaningless w.r.t. the type of musical content under consideration.

# 3 BASIC CONCEPTS

## 3.1 Definitions

As exposed in the previous section, the hypothesis of this work is that the semiotic structure of "conventional" music pieces is built on structural *blocks*, characterized by the content of their musical layers. One of the aim of semiotic structure annotation is therefore to locate the *block boundaries* (with the convention that they are synchronized with the first beat of a bar). We call *size* the dimension of the blocks relative to a *snap* scale proportional to that of the beat (see 3.3).

We call *structural metric pattern*, the underlying high-level organization of the musical content which is the most adequate for representing economically the semiotic level, and we assume that block boundaries rest on the (potentially irregular) *realization* of that structural metric pattern. The annotation task thus consists in jointly inferring the structural metric pattern and its realization.

## 3.2 Musical information layers

Even though this is a simplified view of reality, we consider that a piece of music is characterized by *4 main reference properties*, potentially evolving over time[2] :

- intensity (amplitude / sound level)
- tonality/modality (reference key and scale)
- tempo (speed / pace of the piece)
- timbre (instrumentation / audio texture)

We also consider that a piece of music shows *4 main levels of temporal organization* :

- rhythm (relative duration and accentuation of notes)
- melody (pitch intervals between successive notes)
- harmony (chord progression)
- lyrics (linguistic content and, in particular, rhymes)

[2]In previous work, we identified 3 reference properties only, but we consider now that *intensity* should also be part of the list.

These levels of description form *8 musical layers* [3].

Because of their cyclic properties in conventional music, the levels of temporal organization are central to the *determination* of block boundaries, in our approach. Indeed, as explained in section 4.1, we assume that block boundaries coincide with the convergence of cyclic behaviors taking place simultaneously in the 4 levels of temporal organization.

On the opposite, blocks may globally differ in terms of intensity, tonality, tempo or timbre but these properties may happen to change within a block without corresponding to a structural boundary.

### 3.3 Block size

A primary property of blocks is their size, which we describe in a custom unit that we call *snap*, and which is defined as the number of times a listener would snap his fingers to accompany the music, at a rate which is as close as possible to 1 bps (beat per second). As opposed to the beat (which is a compositional notion), the snap is a perceptual unit.

Although we may come to consider the blocks from a variety of perspectives during their identification, their ultimate description will be their size in snaps. The definition of the snap requires further consolidation, since a tempo-invariant unit would be desirable. However, an evolution of the definition of the snap would not affect the structural segmentation *per se*, as the snap is only a measure of the block size.

### 3.4 Structural metric pattern

A fundamental assumption of the proposed method is based on the hypothesis that the semiotic structure can be described in reference to a *structural metric pattern*, *i.e.* a prototypical partition of the beat or the snap scale. As an example, a very common structural metric pattern is the repetition of blocks of 16 snaps (structural pulsation period $\Psi = 16$).

The high-level structure of the music piece is governed by the structural meter but actual semiotic blocks result from the *realization* of the structural meter and this realization may lead to blocks of *irregular* size. For example, even if the structural period of a piece is equal to 16, the size of some blocks may deviate from the prototypical value (for instance, 18). We develop further the fact that, in a large number of cases, irregular blocks can be *reduced* to regular *stems* that conform to the structural metric pattern.

The structural metric pattern is analogous to the bar, but operates at a higher level : whereas the bar is the organizational entity of low-level elements such as beats and notes, the structural metric pattern governs the organization of mid-level elements (bars, cells, phases, etc...).

## 4 ANNOTATION CRITERIA AND NOTATION

### 4.1 Detection of cycles (syntagmatic analysis)

In conventional music, the various temporal organization layers tend to show (quasi-)cyclic behaviors, which we define as the recurrent return of the considered layer to some specific state or set of states[4]. For instance, rhythmic patterns generally show a short-term recurrence which participates to the mid-level organization of the music piece, melodies tend to return to tonic or to exhibit particular intervals (depending on the piece), specific chords sequences conclude harmonic progressions (cadences), etc...

We consider that, in conventional music pieces, there exist time instants for which the 4 levels of temporal organization exhibit some *phase convergence* towards their respective ends of cycles, which creates identifiable *cues* of the piece structure. In other words, block boundaries should correspond to some form of recurrent convergence of all levels of temporal organization.

These instants of convergence take very versatile forms, as they can be signaled in the music content by very diverse combination of *structuring cues*, such as a particular rhythmic pattern combined with the return to a specific note or chord, the completion of a system of rhymes in the lyrics the conclusion of a *carrure* and a recurrent sound effect...

Even though these cues and their combinations are partly conventional (at least within a particular music genre), they generally vary from one piece to another and their identification is part of the empirical analysis conducted by the annotator.

In our approach, cyclicity plays a central role for identifying structural blocks through the 2 ensuing properties :

1. *iterability* : structural blocks can be looped to yield a consistent (larger) musical stream
2. *suppressibility* : structural blocks can be skipped in the music piece without creating the perception of a discontinuity in the remaining musical stream

[3] These layers may not all be active simultaneously and some additional layers may be observed in some music pieces.

[4] Note that *cyclic* does not necessarily mean periodic, the latter being a stronger property. For example, the zero-crossing of a sequence of values form a set of cycles which may not be periodic.

Indeed, if one thinks of a periodic signal, each period can be repeated indefinitely and can be removed from the signal without disrupting seriously the organization of the remaining signal. This generalizes conceptually to quasi-cyclic processes, as defined above.

The property of cyclicity gives a founded ground for the syntagmatic definition of structural blocks. It establishes more clearly the criterion formerly based on the preservation of "musical consistency" [12] and also brings additional substance to the concept of Constitutive Solid Loop [11].

The listener's ability to identify iterable and suppressible segments in the music piece is a key point in the proposed analysis and it does not require the annotator to be able to express in musicological terms the actual properties of the structuring cues.

When necessary, the analysis can be complemented by an explicit designation of the structuring cues, but attention must be paid that these cues should not be expected to be univocally associated to blocks boundaries : all structuring cues are not systematically observed at all segment borders and some cues can also be observed within block boundaries.

## 4.2 Detection of similarities (paradigmatic analysis)

The identification of actual block boundaries is further (or, in practice, simultaneously) carried out by performing paradigmatic analysis on the musical content, for reinforcing and disambiguating the set of candidate borders hinted by the detection of cyclic segments.

It consists in searching for "repeating" patterns across the musical content, which are identical, similar or, more generally speaking, *easy to explain economically* relative to one another (for instance, transposition, level of instrumental support, addition of a melodic motif, insertion of a musical segment, ...).

As for the syntagmatic analysis of section 4.1, the locations of such paradigms do not coincide univocally with block boundaries : they only constitute additional cues of such boundaries.

Note that the paradigmatic analysis performed at this stage calls for similar processes to those that are needed for labeling the segments. However, whereas the labeling stage requires the determination of a global system of contrasts between segments, the extraction of paradigmatic structural cues simply requires *pairwise* comparisons of musical segments for the only purpose of identifying and locating candidate blocks.

## 4.3 Regularity and reduction

For many conventional music pieces, it can be assumed that a majority of blocks within the piece have a comparable size in snaps, hence corresponding to some structural pulsation period ($\Psi$). Blocks whose size is equal to the structural pulsation period are called *regular* blocks.

Some blocks have a smaller size than $\Psi$, which can generally be interpreted as corresponding to a shortened realization of a regular block. This is especially true for half-size target segments, which can often be matched with the first or second half of a regular block observed somewhere else in the piece. Alternatively such blocks may be considered as a half realization of the structural metric (this is often the case for pre-chorus and bridges).

In a significant number of cases, blocks are longer than the structural period. However, in these cases, they can often be reduced into a *stem* of size $\Psi$ and an *affix*. An affix is a subset of snaps which can be viewed as having been inserted into a (regular) stem and affixes are therefore suppressible from the original block (but not necessarily iterable), *i.e.* the stem forms, on its own, an admissible block. If the insertion of the affix takes place at the beginning (resp. at the end) of the block, it is called a prefix (resp. suffix).

Affixes are particularly easy to identify and locate within a block when there exist, somewhere else in the song, another block which corresponds to the realization of the stem alone. But sometimes, the stem has to be hypothesized based on more subtle considerations, because it is not attested alone in the piece (but, for instance, with a different affix).

Frequent examples of suffixes are observed when for instance a block is extended by lengthening the last snap over 2 more snaps (resulting in some form of break), by doubling the duration values of the notes on the last 2 snaps of the block or by repeating the last 4 snaps twice (thus rendering an insistence effect). Affixes within blocks can be more tricky to detect, and may take versatile forms, for instance the repetition of a $p$-snap segment, a tonal excursion of a few snap or a segment with totally different properties from the rest of the block.

By convention, prefixes and suffixes should be of maximum size equal to half of that of the block (preferably strictly less) and they should not alter the harmonic *valence* of the block, *i.e.* the harmonic properties at the block boundaries.

### 4.4 Structural metric pattern notation

To describe the structural metric pattern, we use the following notation :

| | |
|---|---|
| $n$ | a constant stem size of $n$ snaps throughout the piece |
| $\{n_1, n_2\}$ | 2 stem sizes in the piece, $n_1$ and $n_2$, occurring in any order but in decreasing frequency (can be generalized to more than 2 values) |
| $(n_1, n_2)$ | a systematic alternation of stem sizes $n_1$ and $n_2$, starting with $n_1$ (can be generalized to more than 2 values) |

These notations are superscripted with a star ($n^*$, $\{n1, n2\}^*$, etc...), if the piece contains only within-blocks irregularities, or very few short blocks considered by the annotator as non-representative of the dominant structure of the piece (in particular, in intros, outros, re-intros, etc...). If relevant, the annotator can combine further the notations, for instance $\{16, (12, 8)\}$, but these needs are quite exceptional...

In conventional pop music, the most common segmental structure is $m$ x $16^*$ ($m$ being the number of blocks, which is itself usually close to 16), but pieces from the genre *blues* have usually block sizes based on 24 snaps. More complex patterns such as $\{16, 12\}$, $(16, 8)$ or $(16, 16, 8)$ happen to be observed.

### 4.5 Block size notation

Following are the corresponding notation conventions which we use to designate the size of (realized) blocks, in reference to a structural pulsation period of $n$ snaps:

| | |
|---|---|
| $[p + n]$ | Insertion of a $p$-snap prefix before stem |
| $[n\&p]$ | Insertion of a $p$-snap infix (somewhere) inside stem |
| $[p - n]$ | Omission of $p$ snaps at the end of stem |
| $[-p + n]$ | Omission of $p$ snaps at the beginning of stem |
| $[n\backslash p]$ | Omission of $p$ snaps (somewhere) inside stem |
| $[n/2]$ | Half-size block (undetermined place of missing half) |
| $[x]$ | Undeterminable size (usually owing to a lack of snap) |

Sometimes, two structural blocks may overlap over $p$ snaps, which we call block *tiling*. This is the case when the realization of a new block starts while the previous blocks is still $p$ snaps before its final boundary and continues in the meantime (for instance, in canons). It is also the case when some snaps function simultaneously as the end of a given block and the beginning of the next one. The notation convention for tiling situations is : $[n - p\ [p]\ p + n]$.

Note that the internal structure of blocks could be further specified by decomposing the block size into sub-blocks according to paradigmatic properties within the block (for instance 4x4 as the internal structure of a size 16 block), but this goes beyond the scope of the current paper.

## 5 GENERAL METHODOLOGY

### 5.1 Annotation process

Based on the notions introduced in the previous section, the annotation of a music piece $X$ can be understood as an (empirical) joint estimation task, namely the determination of :

- The most likely structural metric pattern ($M$) for the piece.
- The most likely decomposition of the piece into a set of blocks (S), *i.e.* the realization of $M$.

In practice, the annotator proceeds iteratively as follows :

1. *hypothesize* a structural period $\Psi$, or (more generally) a structural metric pattern $M$ from the listening of X.
2. *(attempt to) decompose* $X$ into blocks following $M$, by introducing, if and only if necessary, irregularities (affixes, irregular blocks) so as to satisfy cyclicity of blocks and to maximize similarities across blocks (resp. sections 4.1 and 4.2).
3. *consider* possible alternatives to $\Psi$ or $M$.
4. *if* such alternative(s) seem to be worth considering, *return to step 2* and test the new hypothesis.

The understanding of step 2 is crucial to the proposed methodology : at that stage, the annotator is actually trying to estimate the realization of $M$ via the minimization of the necessary distortion that $M$ should undergo to make it match the properties of the actual musical content of $X$. Ultimately, among various hypotheses for $M$ and the corresponding decompositions, the annotator retains that which seems globally more economical for describing the semiotic level, *i.e.* the solution which results in a satisfactory compromise between :

- the simplicity and typicality of the structural metric,
- the regularity of the decomposition,
- the non-redundancy of successive blocks,
- the closeness of the structural period(s) to a reference value (currently set to 15 seconds).

## 5.2 Hypothesizing the structural metric pattern

### 5.2.1 *A priori* properties and typical values

Previous work has put forward arguments based on the "Predictive Information Context" (PIC) which suggest that the *a priori* most economical decomposition of a music piece into structural units is based on segments whose typical length would be equal to the square root of the length of the piece.

In an annex to this paper, we propose complementary considerations based on information theory concepts, which strengthen this point.

As a consequence of this property, we assume that structural blocks of a size approximately equal to the square root of the length of the piece happen to be a *reasonable initial assumption* when looking for possible hypothesis of the structural pulsation period. However, the actual analysis of the musical content may lead to a final (*a posteriori*) result which deviates significantly from this starting point.

Numerically, on the basis of an average song length of 240 seconds and a snap value around 1 s, $n = \sqrt{N}$ falls in the range of 15.5 s, which is roughly the typical duration of blocks used so far as target value in our annotations.

With a snap around 1 second, the size of a block will therefore typically be of 16 snaps. Here again, this property should only be considered as *a priori* hypothesis (the one to start with).

From these consideration, a *canonical model* which summarizes all the *a priori* can be laid down : it consists in 16 blocks of 16 snaps of 1 s each. For a given piece, the structural metric pattern and its realization are thus searched as the *minimal deviation from this canonical model*, which enables a structural description compatible with the musical content.

### 5.2.2 Estimating plausible snap ans structural period(s)

By definition, the snap is the multiple of the beat corresponding to a duration as close as possible (in logarithmic scale) to 1 s (in fact, it usually corresponds to the downbeat, but not always). Therefore, identifying the snap is, in general, rather straightforward from the listening of parts of the piece, preferably away from the beginning or the end, which may exhibit particular beat and tempo properties. Depending on the type of bar, admissible intervals for the snap are : [0.71, 1.41] for binary bars and [0.58, 1.73] for ternary ones (for more complex, odd bars, the snap can be unevenly alternating between different numbers of beats).

Once the snap is determined, plausible values of the structural pulsation period(s) are hypothesized by listening to the piece and considering in priority the most salient and steady parts of the piece : typically the chorus (if any), the developments of recurring motifs or phrases, the parts of the piece perceived as relatively homogeneous, etc... From these segments, the annotator can generally infer rapidly one or two plausible values of pulsation period(s), from which he/she will start a more comprehensive analysis of the entire musical material of the piece, looking for particular patterns and locating irregularities.

As a consequence of the central role played by the canonical model, the value of 16 is usually investigated in priority, unless obvious evidences in the musical content directs the annotator towards another hypothesis (for instance, 24 in many pieces of blues).

# 6 A CASE STUDY

Figure 1 illustrates the analysis of song Genre 08 from the RWC database [17]. Structural blocks are depicted both as their span on the x-axis (time in snap) and their height on the y-axis (in log scale). Each block is identified by a distinct roman number.

The duration of the song file is 3'26" (including initial and final silences) and the size of the actual song in snap is 200 (snap duration almost exactly equal to 1 s). The following observations are based on the listening of the music piece.

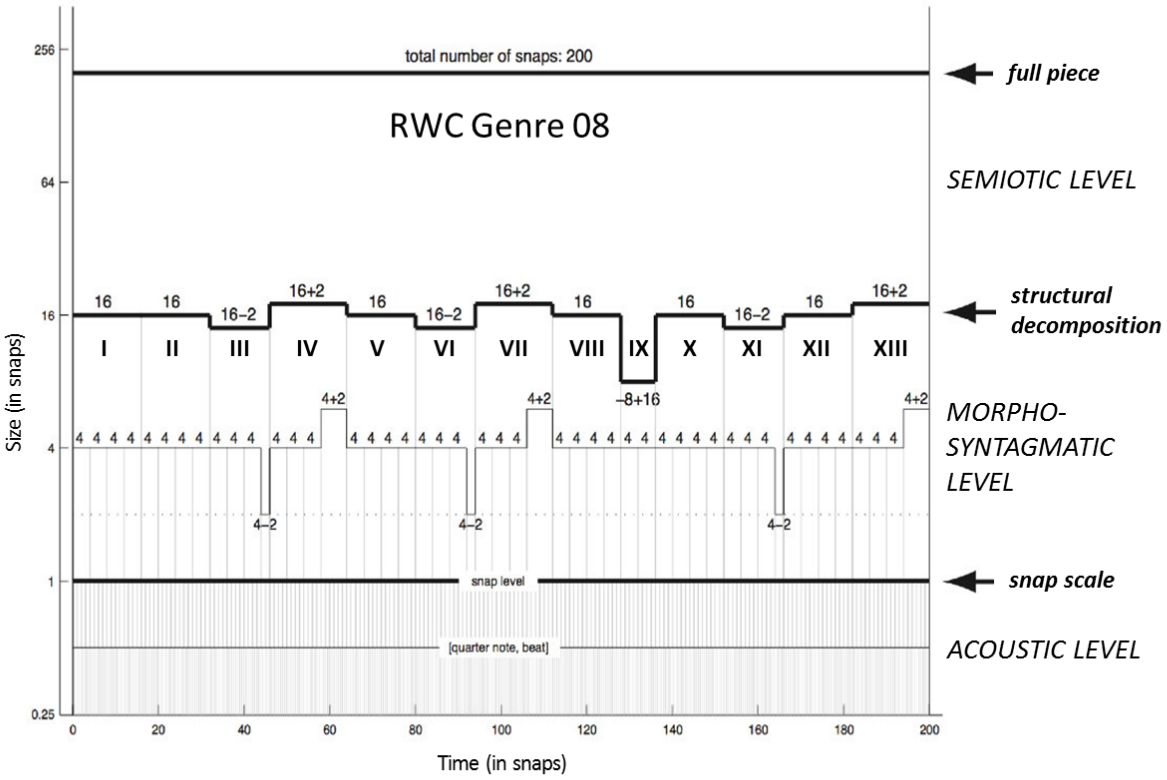


Figure 1: RWC Genre 08 annotation levels

Segments IV, VII, XII and XIII present a clear paradigmatic relationship (they are easy to qualify as the chorus of this piece). Three of them last 18 snaps but XII lasts only 16 snaps and can be considered as the stem on which the three other blocks are built by lengthening the harmonic content over the last 2 snaps.

Segments II, V, X form a second paradigm, with the return to tonic as a clear (conventional) structuring cue. Being of size 16, they are in line with the $\Psi = 16$ hypothesis. An alternative hypothesis would be to consider them as the repetition of 2 almost identical (half-)blocks of 8 snaps, but i) this would need the introduction of a second structural period and ii) no occurrence of such a half-block *alone* is observed in the song and iii) it would split the rhyming pattern of block V.

Segments III, VI and XI constitute a third paradigm. Their raw form amounts for 14 snaps, but they can be described as a 4x4 snap *carrure* of the *abab* type, whose last quarter has been truncated of the last 2 snaps, hence the notation $16 - 2$. This comforts (or at least does not contradict) the hypothesis $\Psi = 16$.

Segments I and IX are very similar, I being an instrumental intro of 16 snaps and IX the second half of I, used as an instrumental bridge (hence the notation $-8 + 16$). Finally, VIII is a solo, which conveniently lasts exactly 16 snaps.

The segmental structure of the piece is therefore considered to be 13 x $16^*$, *i.e.* a basic 16-snap pattern realized 13 times with a few within-block irregularities. Alternative options could have been 25 x $8^*$, but this would introduce much redundancy in the underlying semiotic description, since almost all segments would be observed in systematical pairs, without bringing significantly down the number of irregular segments (only IX would thus become regular). A pattern such as $(16, 14, 18)^*$ could be envisaged given the recurrence of this particular size sequence in II-III-IV and V-VI-VII but the existence of XII as a 16-snap realization of the chorus just in between XI and XIII makes this complicated alternative a non-sustainable option.

# 7 CORPUS DESCRIPTION

## 7.1 RWC Pop set

Part of the available annotations is composed of the 100 Pop songs from the RWC database [17], written and produced for research purposes. Their structural annotations have been released and used last year for the MIREX 2010 evaluation [18] in structural segmentation and they are currently under minor revision.

| | |
|---|---|
| **RWC Pop** | **100 titles** |

## 7.2 Quaero set

The *Quaero set* is composed of 114 titles which have been selected by IRCAM and used in the Quaero project [13] for the evaluation of music structure detection algorithms :

| | | |
|---|---|---|
| Quaero 2009 | Development set | 20 titles |
| Quaero 2009 | Evaluation set | 49 titles |
| Quaero 2010 | Evaluation set | 45 titles |
| **Total** | | **114 titles** |

The average length of songs in this corpus is approximately 4 minutes. A subset of 55 titles contains several pieces from the same artists (The Beatles : 9, Eric Clapton : 7, Pink Floyd : 7, Queen : 7, The Cure : 6, Jedi Mind Tricks : 5, D Angelo : 4, ACDC : 2, Eminem : 2, Madonna : 2, Plastikman : 2, Shack : 2) and the remaining 59 titles correspond to 59 other distinct artists. In this corpus, a large range of music genres is covered but the vast majority of artists are American or English.

## 7.3 Eurovision set

The Eurovision set is currently composed of 124 titles, corresponding to the songs which participated to the semi-finals and/or the final in years 2008, 2009 and 2010, in their studio version (as recorded on the "official" albums) :

| | | |
|---|---|---|
| 2008 (Belgrade) | ref # 5 099921 699726 | 43 titles |
| 2009 (Moscow) | ref # 5 099969 968020 | 42 titles |
| 2010 (Oslo) | ref # 5 099964 171722 | 39 titles |
| **Total** | | **124 titles** |

Eurovision songs have the particularity of being limited to a 3'00" maximum duration by the rules of the contest, and they tend to show other properties (including their structure) influenced by the contests format and its target public. These titles however cover a variety of languages and a diversity of sub-genres within European pop music.

### 7.4 Ongoing effort

At the time of writing this paper, we are finalizing the annotation of the RWC Genre database (100 titles). We have also planned to annotate a new set of 50 titles for the Quaero project and will complete the current effort by an additional set of 12 titles, so as to reach, together with the already achieved annotations, a total of 500 annotated titles by the end of the summer 2011.

### 7.5 Release

All the aforementioned annotations will gradually be made available online before September 2011 at :

http://musicdata.gforge.inria.fr

and a number of them will be accompanied with comments as case studies (such as in section 6), so as to document the annotation method.

## 8 CONCLUSIONS

The work presented in this paper constitutes a contribution towards the general strategic goal of disseminating consistent and re-usable resources for research and development in MIR. It also contributes to the objective of converging towards operational concepts for the description of music structure, through the definition of the *structural metric pattern*, and a consistent annotation procedure. Together with the production of additional resources, our current work direction is to consolidate the connections between music structure description and information theory, so as to encompass a wider range of concepts and, in particular, to integrate several timescales in the structural description.

## References

[1] J. Paulus, M. Muller and A. Klapuri, "Audio-based music structure analysis", *Proceedings of the International Symposium on Music Information Retrieval*, pp. 625–636, 2010.

[2] G. Peeters, "Deriving Musical Structures from Signal Analysis for Music Audio Summary Generation: Sequence and State Approach", *Lecture Notes in Computer Science*, Springer - Verlag, 2004.

[3] S. Abdallah, K. Noland, M. Sandler, M. Casey and C. Rhodes, "Theory and evaluation of a Bayesian music structure extractor", *Proceedings of the International Symposium on Music Information Retrieval*, London, UK, 2005.

[4] M. Goto, "A Chorus Section Detection Method for Musical Audio Signals and Its Application to a Music Listening Station", *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[5] J. Paulus and A. Klapuri, "Music structure analysis by finding repeated parts", *Proceedings of AMCMM,* Santa Barbara, California, USA, 2006.

[6] G.Peeters, "Sequence Representation of Music Structure Using Higher-Order Similarity Matrix and Maximum- Likelihood Approach", *Proceedings of the International Symposium on Music Information Retrieval*, Vienna, Austria, 2007.

[7] M. Levy and M. Sandler. "Structural Segmentation of musical audio by constrained clustering", *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

[8] C. Kelly, "Locating tune changes and providing a semantic labelling of sets of Irish traditional tunes", *Proceedings of the International Symposium on Music Information Retrieval*, pp. 129–134, 2010.

[9] R. Weiss and J. Bello, "Identifying Repeated Patterns in Music Using Sparse Convolutive Non-Negative Matrix Factorization", *Proceedings of the International Symposium on Music Information Retrieval*, pp. 123–128, 2010.

[10] F. Kaiser and T. Sikora, "Music structure discovery in popular music using non-negative matrix factorization", *Proceedings of the International Symposium on Music Information Retrieval*, pp. 429–434, 2010.

[11] G. Peeters and E. Deruty : Is Music Structure Annotation Multi-Dimensional ? A Proposal For Robust Local Music Annotation. *LSAS*, Graz (Austria) 2009.

[12] F. Bimbot, O. Le Blouch, G. Sargent and E. Vincent, " Decomposition into autonomous and comparable blocks : a structural description of music pieces", *Proceedings of the International Symposium on Music Information Retrieval*, pp. 189–194, 2010.

[13] QUAERO Project : http://www.quaero.org

[14] SALAMI Project : http://salami.music.mcgill.ca

[15] B. Snyder, *Music and Memory*, M.I.T. press, 2000.

[16] J.-J. Nattiez, *Musicologie générale et sémiologie*, 1987.

[17] RWC : http://staff.aist.go.jp/m.goto/RWC-MDB

[18] MIREX 2010 : http://www.music-ir.org/mirex/wiki/2010

# 9 ANNEX

Let's consider a song represented as a sequence of discrete elements at a given time-scale $X = \{x_k\}_{1 \leq k \leq N}$ and let's now consider a bi-dimensional organization of $X$ into blocks of size $q$, *i.e.* a $m$ (lines) $\times$ $n$ (columns) matrix representation of $X$ :

$$X = [x_{i,j}]_{1 \leq i,j \leq m,n} \tag{1}$$

with $m = N/n$ and $k = (i-1)n + j$.

Given this structure, the quantity of information needed to index all elements in the matrix requires :

$$I_n = m \log(m) + n \log(n) = \frac{N}{n} \log(\frac{N}{n}) + n \log(n) \tag{2}$$

Thus, the index of each line in the matrix $X$ can be coded with $\log(m)$ bits, and the total number of bits required to index all lines in the matrix is $m \log(m)$ (the same applies for the columns, hence $n \log(n)$). Seeking for the minimum of $I_n$ (by zeroing the derivative of $I_n$ w.r.t. $n$) yields $n = \sqrt{N}$.

Hence, in the absence of any particular knowledge concerning the redundancies in $X$, the most economical way to *index* it bi-dimensionally is to shape it as a square matrix structure.